Large and noisy vs small and reliable: combining 2 types of corpora for adjective valence extraction

Cécile Fabre, Anna Kupść

Large and noisy vs. small and reliable
combining 2 types of corpora for adjective valence extraction

Cécile Fabre and Anna Kupsc
CLLE-ERSS
University of Toulouse and University of Bordeaux
cecile.fabre@univ-tlse2.fr and akupsc@u-bordeaux3.fr

Abstract

This work investigates a possibility of combining two different types of corpora to build a valence lexicon for French adjectives. We complete adjectival frames extracted from a Treebank with statistical cues computed from a large automatically parsed corpus. This experiment shows how linguistic knowledge and large amount of annotated data can be used in a complementary manner.

1. Introduction

Valency lexicons contain subcategorisation information related to every predicate: in general, the number and type of arguments selected by a predicate (for example, by a verb). As such information is highly lexical and language-dependent, it has to be specified separately for each predicate of the language. In addition to the language learning value, valency lexicons are crucial resources for various NLP tasks and applications, such as parsing (Carroll and Fang, 2004), generation (Danlos, 1985), information extraction (Surdeanu et al., 2003) or machine translation (Han et al., 2000). Initially, such resources have been created manually based on linguistic knowledge of human experts, see among others, (Procter, 1978) and (Hornby, 1989) for English, or (Gross, 1975) and (Mel'cuk et al., 1984-1999) for French. Although such lexicons are readable for humans, they cannot be directly used in computer applications. For example, the best known French valency lexicon of (Gross, 1975) is coded in tables which contain syntactic and semantic properties of predicates. However, the information in the tables is not always stated explicitly and has to be inferred from other properties, which complicates the automatic conversion process, see (Gardent et al., 2006) for details. Another issue related to the existing hand-crafted valency lexicons comes from their coverage as they are not always well-adjusted to contemporary texts. Recent developments in corpus linguistics provided a wide range of methods and resources which allow for creating valency lexicons well-suited for NLP tasks in various languages, see especially (Frank et al., 2002), (Preiss et al., 2007) for methods based on syntactically annotated corpora.

The majority of valence resources has been created for verbs, and much less attention has been paid to specifying valency of other predicates, such as nouns or adjectives. For French, for instance, the two available valency lexicons for adjectives, (Gross, 1975) and (Picabia, 1978), exist only on paper and have not been adapted to automatic processing. In this paper, we describe a method which allows us to create a valency lexicon of French adjectives, adjusted to NLP applications.

Our approach is corpus-based, and the lexicon is automatically extracted, but it combines two different types of corpora. On one hand, we use a relatively small (1 million words) corpus which has been manually revised and enriched with syntactic and functional annotations for major constituents. On the other hand, we have a large (200 million words) corpus automatically parsed, with no subsequent human validation, where the texts have been annotated with dependency relations. In none of the two corpora is the argument/adjunct distinction specified for dependents of adjectives. Our method consists in identifying adjective's arguments exploiting and combining properties of the two corpora: linguistic cues and frequency measures.

The organisation of the paper is as follows. First, we briefly present general properties of French adjectives and issues related to adjective valency. The next two sections describe extraction techniques specific to the two types of corpora. In Section 5, we discuss a method for refining results by adopting a less rigid argument/adjunct distinction. Section 6 concludes the paper and presents perspectives on our future research.

2. Properties of French adjectives

2.1. Types of arguments

In French, complements of adjectives can be realised by three main categories: prepositional phrases (PP), subordinate clauses (Ssub) or infinitival phrases (VPinf).

(1) sûr [PP de son retour] / [Ssub qu'il reviendra] / [VPinf de revenir]
sure of his return that-he will-come-back to come-back
`sure of his return / that he will come back / to come back'

Nominal phrases (NP), on the other hand, can serve only as the subject of an adjective[1]. We adopt the notion of the subject of an adjective both to predicative uses, (2)-(3), where the adjective is a predicate on its own, and to attributive uses, (4), where the adjective modifies a noun that becomes its semantic argument and thus can be considered its semantic subject. Note that in addition to NP, the subject of an adjective can be expressed by Ssub or VPinf, (3).

(2) [$_{NP}$ La maison] est grande. (predicative)
the house is big
`The house is big.'

(3) Jacques trouve inévitable [$_{Ssub}$ qu'elle chante] / [$_{VPinf}$ d'écouter sa chanson].
Jacques finds unavoidable that-she sings to listen her song
`Jacques finds it unavoidable that she sings / to listen to her song.'

(4) Je vois une grande [$_{N}$ maison]. (attributive)
I see a big house
`I see a big house.'

2.2. Specificity of adjectival valence

Although the repertoire of syntactic phrases which can appear as arguments of adjectives is very limited, specifying valence of an adjective can be quite difficult.

First, traditional linguistic tests which help to separate arguments from adjuncts are less reliable than for verbs. For example, one of the strongest criteria used for verbs, the obligatory presence of an argument, is in most cases inapplicable to adjectives as surface realization of a complement is often optional, (5). In fact, (Noailly, 1999) mentions just a few adjectives, such as *enclin* `inclined', *exempt* `exempted' or *désireux* `desirous', among those for which a complement is obligatory. Similarly, results of other `argumenthood' tests, e.g., topicalisation or pronominalisation, are in general less suitable than for verbs, cf. (Picabia, 1978: ch.3).

(5) Paul est amoureux (de sa voisine).
Paul is in love of his neighbour
`Paul is in love (with his neighbour).'

Second, an alternative realization of a PP complement of an adjective is much more common than for verbs. For adjectives, several distinct prepositions may introduce the same semantic argument, see (6), cited after Picabia (1978:85). For verbs, if various prepositions are possible, they normally have to belong to the same semantic class, e.g., the verb *habiter* `to live' accepts various PP complements (*dans* `in', *sous* `under', *à* `in/at', etc.) but they all form a uniform semantic group of locative prepositions. Adjectives seem to be more liberal in this respect as it is difficult to provide a common semantic class which can group *avec* `with' and *envers* `towards' in (6).

(6) Jean est aimable envers / avec Marie.
Jean is pleasant towards with Mary
`Jean is nice towards/with Mary.'

Finally, similarly to verbs, adjectives may participate in many syntactic constructions, for example comparatives or impersonals. It is essential to distinguish components of such high-level constructions from arguments of adjectives. Unlike valency, which depends on individual properties of an adjective, components of productive constructions are much less sensitive to specific adjectives. For example, PP in (7) is part of the superlative construction and it is not required by the adjective itself: *beau* `beautiful' could be replaced by almost any other adjective.

(7) le plus beau [$_{PP}$ de la terre]
the most beautiful of the earth
`the most beautiful on earth'

The above properties of adjectives make valence identification rather challenging. In this paper, due to two different types of corpora, we approach the issue from two different perspectives. On one hand, due to rigid syntactic annotations in the Treebank and linguistic knowledge, we aim at separating valency from components of high-level constructions. On the other hand, large amount of data in the other corpus allow us to adopt frequency tests to detect arguments and verify their variability. The next two sections provide a description of the two techniques.

3. Extracting frames from the treebank

3.1. Treebank

As mentioned above, in the first step, we explore a relatively small (1 million words) but richly annotated corpus. We use the Treebank of Paris 7 (Abeillé et al., 2003), a corpus consisting of 4 years of *Le Monde*, a French daily newspaper. The text has been segmented into words and phrases and then linguistically annotated. The initial annotation was done automatically but then it has been validated by human experts. Linguistic information in the corpus concerns words or lexical compounds, indicating the category, morphological properties and the lemma, as well as phrases, specifying the category of a constituent and a grammatical function. A sample of corpus annotations for the sentence *Paul est fier de ses enfants* `Paul is proud of his children' is given in Fig. 1.

```
<SENT>
 <NP fct="SUJ">
     <w cat="N" m="N-P-ms" lemma="Paul">Paul</w>
 </NP>
 <VN>
     <w cat="V" m="V--P3s" lemma="être">est</w>
 </VN>
 <AP fct="ATS">
     <w cat="A" m="A-ms" lemma="fier">fier</w>
     <PP>
             <w cat="P" m="P" lemma="de">de</w>
             <NP>
                     <w cat="Det" m="D-poss-pl" lemma="se">ses</w>
                     <w cat="N" m="N-C-mp" lemma="enfant">enfants</w>
             </NP>
     </PP>
 </AP>
</SENT>
```

Figure 1: *Paul est fier de ses enfants* `Paul is proud of his children'

As can be seen from this example, the constituent structure is rather flat: there is no VP and all dependents of the verb (or more general, the verbal nucleus, VN) are related to it only via grammatical functions: both the subject (SUJ) and the subject complement (ATS) in the example form independent phrases. Note that functions are specified only for verb dependents: the PP complement of the adjective *fier* `proud' is structurally embedded within the AP but its function with respect to the adjective is not indicated in the corpus. Similarly, the subject of the predicative adjective is not provided either: the sentential subject, i.e., NP, *Paul,* is shared between the copula (*est* `is') and the adjective but this link is not specified in the treebank.

As Fig. 1 shows, adjectival valence is not directly indicated in the corpus. In order to obtain it from the treebank, we combine linguistic knowledge with corpus annotations.

3.2. Extraction method

Our extraction method is guided by linguistic cues applied to treebank annotations. We focus on AP constituents and restrict the types of phrases that can appear as arguments of an adjective to categories indicated in sec. 2.1, both for complements and the subject. Our main goal is to distinguish regular adjectival constructions from valency components.

3.2.1. Arguments

In the Treebank, predicative adjectives are direct arguments of a verb and they are assigned grammatical functions: a subject complement (ATS) or an object complement (ATO), i.e., a predicate referring either to the sentential subject (2) or to the direct object (8).

(8) [$_{NP}$ Jacques] trouve [$_{AP}$ inévitable] [$_{Ssub}$ qu'elle chante].
SUJ Jacques finds ATO unavoidable OBJ that-she sings
`Jacques finds it unavoidable that she sings.'

In such cases, the subject of the adjective can be easily identified as it is indicated by the grammatical function of another argument of the verb: SUJ for ATS, and OBJ for ATO adjectives, as in (8).

Adjectives may appear also in impersonal constructions with an accompanying Ssub or VPinf, (9). The status of the propositional components in (9) is different from those in (10), as illustrated also by the corpus annotations. The crucial difference is that Ssub or VPinf in (9) can be preposed to become the sentential subject, whereas this is not possible in (10).

(9) Il est [$_{AP}$ agréable] [$_{Ssub}$ qu'il fasse beau] / [$_{VPinf}$ de sortir].
it is ATS nice OBJ that-it makes beautiful OBJ to go out
`It's nice that the weather is good / to go out.'

(10) Paul est [$_{AP}$ heureux [$_{Ssub}$ qu'il fasse beau] / [$_{VPinf}$ de sortir]].
Paul is ATS happy that-it makes beautiful to go out
`Paul is happy that the weather is good / to go out.'

In (10), Ssub or VPinf is embedded within AP, unlike in (9). The propositional constituents in (9) become the extraposed subject of the adjective, i.e., in impersonal constructions (the subject is expressed by pronouns *il* or *ce*), OBJ-phrase is in fact the subject of ATS adjective. On the other hand, if no construction-related elements are present (sec. 3.2.2), the subordinate components, as in (10), are treated as complements of the adjective.

French clitics are always attached to a verb but they can replace dependents of other predicates as well. Although clitics often pronominalise arguments, they can refer to adjuncts as well, for instance to locative phrases. In the corpus, clitics are direct dependents of a verb and they are assigned a function. In copular predicative constructions, (11), as the copula itself does not have a clitic argument, if the function assigned to the clitic indicates an argument (A-OBJ in (11)), it must be an argument of the predicative adjective. The category of the argument is restored based on the form of the clitic and its function.

(11) Paul [$_{VN}$ y est] [$_{AP}$ favorable].
Paul A-OBJ to-it is ATS in favour
`Paul is in favour of it.'

3.2.2. Non-arguments

Constituents which regularly appear in well-defined syntactic constructions are not related to a specific adjective and do not belong to its valence list. We filter out such PP, VPinf or Ssub by linguistic cues.

In comparative constructions, an adjective is often accompanied by a phrase annotated in the corpus as an internal PP or Ssub component of AP, (11). Note that in such sentences, in contrast to (10), the adjective additionally appears with a comparative adverb, *plus* `more', *moins* `less', *autant* `as much as', etc. Therefore, we exclude the embedded constituent from the list of adjective arguments, unlike in (10) where there is no adverb.

(12) La réunion était [$_{AP}$ plus intéressante [$_{Ssub}$ que je ne pensais]].
the meeting was ATS more interesting that I NOT thought
`The meeting was more interesting than I thought.'

(13)-(14) illustrate another type of productive constructions where the embedded constituent of AP is not an argument of the adjective. Again, the presence of intensifier adverbs, such as *si* `so', *trop* `too', *tellement* `so much', etc., is decisive for the status of Ssub or VPinf constituent within AP.

(13) Paul est [$_{AP}$ si heureux [$_{Ssub}$ qu'il saute de joie]].
Paul is ATS so happy that-he jumps of joy
`Paul is so happy that he jumps out of joy.'

(14) Cette histoire est [$_{AP}$ trop belle [$_{VPinf}$ pour être vraie]].
this story is ATS too beautiful for be true
`This story is too good to be true.'

3.2.3. Lexicon of prepositions

Apart from comparatives, prepositional phrases do not appear in adjectival constructions. Therefore, no other linguistic observations can help us to specify the status of PPs in APs. In particular, there is no general rule which would permit to distinguish a PP complement of an adjective from a PP in the restructured complex NP subject, cf. (Meydan, 1999). Instead, we use PrepLex (Fort and Guillaume, 2007), a lexicon which specifies for each preposition whether it can introduce an argument of a verb. We adopt it to filter out PPs which cannot be complements of an adjective. Additionally, for adjectives, we exclude one preposition, *comme* `as', from the list of argumental prepositions, as in APs it is used only in comparative constructions.

3.3. Results

The presented method results in a list of 2153 adjectives and discovers 40 frames. Each frame indicates the category of the subject and complements (if any). If no complement and no propositional subject have been found for an adjective in the corpus, we assume that its valency list contains only the NP subject. (We refer to this frame as basic.) The majority of adjectives (1849) were found only with the basic frame whereas 304 had a different subcategorisation pattern. Table 1 presents 23 extracted frames which appeared at least twice in the Treebank, their frequency counts and the number of adjectives with which they were found.

Frame	Frequency	#adjectives
SUJ:NP (basic)	15485	2087
SUJ:NP|P-OBJ:PP[à]	278	81
SUJ:NP|P-OBJ:PP[de]	204	94
SUJ:NP|P-OBJ:VPinf[de]	83	44
SUJ:VPinf[de]	66	29
SUJ:NP|P-OBJ:VPinf[à]	53	16
SUJ:NP|P-OBJ:PP[pour]	35	29
SUJ:NP|P-OBJ:PP[en]	30	23
SUJ:NP|P-OBJ:VPinf[pour]	24	6
SUJ:NP|P-OBJ:PP[dans]	22	14
SUJ:SsubI[que]	18	11
SUJ:NP|OBJ:Ssub[que]	18	4
SUJ:NP|P-OBJ:PP[par]	13	12
SUJ:NP|OBJ:SsubI[que]	12	3
SUJ:NP|P-OBJ:PP[sur]	11	11
SUJ:NP|P-OBJ:PP[avec]	9	6
SUJ:NP|P-OBJ:PP[loc]	8	8
SUJ:NP|P-OBJ:PP[entre]	5	3
SUJ:SsubS[que]	6	5
SUJ:NP|P-OBJ:PP[chez]	4	3
SUJ:NP|P-OBJ:PP[depuis]	3	3
SUJ:VPinf[de]|P-OBJ:PP[à]	3	3
SUJ:NP|P-OBJ:PP[après]	2	2

Table 1: Extracted frames with their frequency and the number of adjectival entries in which they appear. Abbreviations: functions: SUJ – subject, P-OBJ – PP or VPinf object, OBJ – object without an introducing element; categories: PP – prepositional phrase, Ssub – a subordinate clause, either in subjunctive (SsubS) or indicative (SsubI) mode, VPinf – an infinitive clause.

In order to get a better grip of the obtained results, we provided a brief examination of the extracted frames. This investigation revealed a few issues. First, due to imperfect or insufficient treebank annotations, the data is not totally reliable. In particular, subjectless or embedded impersonal constructions, (15), are unrecognized in the corpus. The absence of the imminent impersonal subject yields to incorrect or missed argument assignment: VPinf is either misinterpreted as the object of the adjective or is not taken into consideration at all.

(15) (Il pourrait être) impossible [$_{\text{VPinf}}$ d'ignorer les liaisons transatlantiques].
it could be impossible OBJ to ignore the relations transatlantic
`It would be impossible to ignore transatlantic relations.'

Second, Preplex does not allow us to efficiently separate real PP-arguments from adjuncts. All argumental prepositions (i.e., P which can introduce an argument) listed in the lexicon are ambiguous as they can be used in PP-adjuncts as well. For instance, *de* `of' in (1) indicates a PP complement, whereas in (7) the same preposition introduces a PP which is not subcategorized for. Therefore, each preposition has to be considered individually with its adjectival context. Moreover, Preplex has been created for verbs. It should be verified to which extent the list is valid also for adjectives. For example, *comme* `as', listed among argumental prepositions for verbs, had to be moved to a non-argumental list for adjectives.

Finally, due to the corpus size, certain adjective-frame realisations are missing. For example, *plausible* `plausible' has been found only with the basic frame in the corpus.

In order to improve the quality of the adjectival lexicon and extend its coverage, we complement the presented results with data from a much larger corpus.

4. Using a large automatically annotated corpus

In the second step, our method extracts frames by using a much larger corpus and applying statistical methods with two objectives in mind. First, we aim at improving the extraction that has been performed in the previous step. The frames discovered in the treebank are now considered candidate frames that will be either corroborated or invalidated by new corpus data. Second, we use additional corpus data to find new frames by examining occurrences of hundreds of new adjectives.

We have chosen to focus on verification of PP and VPinf complements as they turned out to be the most problematic in the previous step. More precisely, we now consider the set of lexicalised frames that have been discovered in the treebank, i.e., instances of candidate frames, containing a PP or VPinf argument, with a specific adjective, for example:

(16) sûr VPinf [*de*]
'sure of'

(17) aimable PP[*envers*]
'pleasant with'

Obviously the set of candidate frames could simply be manually validated, since the number of different lexicalised frames is not insurmountable (if we put aside adjectives with the basic SUJ:NP frame, cf. Table 1). However, the idea is to use them as a first data set to evaluate statistical tests we have applied to filter candidate frames. In particular, we want to make sure that the tests are relevant and can be in turn applied to help identifying new lexicalised frames in a large corpus.

4.1. Extracting patterns

In order to obtain information about adjectives, we examine the output of a French syntactic parser, Syntex (Bourigault 2007), applied to a large corpus. The corpus is composed of texts from *Le Monde* (from now on called the LM corpus), the same French daily newspaper which served to build the Treebank, but is much bigger: the LM corpus contains 200 million words. Syntex provides dependency information using a combination of heuristics and statistical cues. The parser does not always build a complete syntactic structure: some units can be left unattached when information is insufficient to resolve dependency relations.

We extract all adjectives with prepositional dependents, i.e., prepositions introducing either a noun phrase or an infinitive clause. Syntex identifies four types of relations (functions) related to adjectives. The ADJ relation indicates attributive uses, cf. (4). Predicative adjectives, the subject (2) or object complement (3), are marked as ATTS and ATTO respectively. Appositive adjectives are indicated as APPOS. Example (18) shows which information, morphosyntactic and dependency, is obtained for each adjective.

(18) Les fortes pluies consécutives aux deux tempêtes
'The heavy rains resulting from two storms'
AdjFP|consécutif|consécutives|4|ADJ;3|PREP;5

Every extracted adjective contains information about its governor (ADJ;3 means that the adjective is connected to its governor – 3rd token of the current sentence – via the relation ADJ) and its dependent (PREP;5 means that the adjective governs the 5th token via the relation PREP). Additionally, the inflected form and the lemma are specified.

In 29% of the cases, the parser does not have enough cues to identify the governor of the adjective. In such cases, the relation is left unresolved and a NOGOV relation is assigned to the adjective. In particular, Syntex does not attach appositive adjectives when they occur in sentence initial positions, as in example (19).

(19) Soucieuse d'affirmer son indépendance, la banque centrale refusa de céder
'Eager to assert its independence, the central bank refused to give in'
AdjFS|soucieuse|soucieux|1||PREP;2
(no mention of a governor => NOGOV)

Just like in the first step, sec. 3.2, we use heuristics to filter out PP or VPinf elements of syntactic constructions such as superlatives (*le plus* <adj> *des* 'the most <adj> among') or intensifiers (*assez* <adj> 'enough to').

A sample of information that we get from Syntex parses is presented in Table 2. Inflected word forms are reduced to their lemmas. We keep the following information: form and category (N or Vinf) of the dependent, type of the relation, form of the subject (cf. section 4.3.1.). In what follows the resulting triplets ADJ + PREP + DEP-CAT are called patterns.

ADJ	PREP	DEP	DEP-CAT	REL	SUBJ
consécutif 'resulting'	à 'from'	tempête 'tempest'	N	ADJ	pluie 'rain'
enclin 'inclined'	à 'to'	consacrer 'devote'	VINF	ATTS	il 'he'
insensible 'insensitive'	à 'to'	régulation 'regulation'	N	ATTO	système 'system'
soucieux 'anxious'	de 'to'	défendre 'defend'	VINF	APPOS	-
soucieux 'anxious'	de 'to'	affirmer 'assert'	VINF	NOGOV	-

Table 2 : Patterns extracted from the LM corpus

As expected, we get much more data from the large LM corpus than from the Treebank. Table 3 shows that the number of adjective types for which we get dependency information is multiplied by 8, the number of patterns of the form ADJ PREP (VPinf|PP) is multiplied by 18.

	Treebank	**Le Monde**
Nb of adjectives (types)	304	2684
Patterns *PREP* (VINF | N)	26	136
Patterns *ADJ PREP* (VINF |N)	369	6778

Table 3: Quantitative comparison of the two data sets

On the other hand, however, these data are noisy. Automatic parsing is still a challenging task. Syntex has obtained very good evaluation results at the French parsing evaluation

campaign that took place in 2004 (winning the first rank on most corpora) (Paroubek et al., 2007), yet precision figures are ranging from 0.75 to 0.80. This means that if we want to deal with such data, we have to be aware of this error rate, and find solutions to make up for it. Errors can be generated by the word segmentation module and by tagging or parsing programs. We found two main categories of problems: many nouns are wrongly tagged as adjectives and many PPs are attached to an adjective whereas they should be linked to the noun that the adjective modifies.

The obtained data cannot be manually revised, due to their abundance. Instead, our approach consists in relying on the amount of data and in limiting errors by applying simple thresholds which allow us to exclude rare and potentially ill-formed configurations. The next subsections describe these filtering procedures, focusing first on the evaluation of candidate frames from the Treebank (section 4.3) and then turning to the extraction of new frames from the LM corpus (section 4.4).

4.2. Filtering the treebank candidate frames

4.2.1. Impersonal constructions

As mentioned in 3.2.1., VPinf[*de*] in predicative adjectival constructions is ambiguous: in impersonal constructions VPinf is in fact the postposed sentential subject, whereas in personal sentences it is a true VPinf complement. Disambiguating the two cases is a quite straightforward task given a large corpus. In order to identify impersonal constructions, we simply check the form of the subject. In the parser output, information regarding the type of the subject pronoun (personal or impersonal) is not available. We approximate this information by verifying the form of the subject. In French, an impersonal subject can be expressed in two ways: *ce* (and its variants *c'* and *cela*) or *il*, which is ambiguous, corresponding either to a personal (=*he*) or an impersonal form. For each adjective, we calculate the proportion of *ce* or *il* pronouns in the subject position when the parser has identified the ATTS relation.

Impers(adj) = number of *ce*, *c'*, *cela*, *il* pronouns / total number of ATTS relation

We apply this measure to 40 ADJ+VPinf[*de*] constructions that have been extracted from the treebank. We get highly contrasting results, since all the patterns but one get either very high or very low values. We obtain two clear-cut subsets of data:

27 constructions combine with more than 90% of `impersonal' pronouns:

(20) Il est absurde, acceptable, anormal … de + VPinf
It is absurd, acceptable, abnormal … to + VPinf

12 constructions combine with less than 15% of `impersonal' pronouns:

(21) Il est capable, conscient, content … de + VPinf
He is capable, conscious, happy … to + VPinf

This measure enables us to successfully distinguish personal from impersonal constructions. VPinf which appear in patterns with a high percentage of `impersonal' pronouns is considered the subject, whereas VPinf found in patterns with the low percentage of `impersonal' pronouns are treated as a complement.

Only one pattern has an intermediate value and cannot be categorized: *nouveau* VPinf [*de*] (40% of 'impersonal pronouns'). When we look at the contexts, this pattern corresponds mostly to an impersonal construction (*Il n'est pas nouveau de* + VPinf 'It is not new to + VPinf'), but it also matches another kind of construction, (22), where VPinf belongs to the subject (NP headed by an abstract noun) but is placed after the adjective.

(22) L'idée n'est pas nouvelle de + VPinf

the-idea neg-is not new to VPinf

Only the postposed VPinf subject in impersonal constructions must be taken into account in the valence of the adjective.

4.2.2. Valency tests for adjectives

Finding indicators to make distinction between arguments and adjuncts is a difficult task. With respect to verb complementation, there are many linguistic tests but in general they are not conclusive to handle the distinction. A thorough examination of French verb complementation (Bonami, 1999) leads to very specific conclusions, namely that: 1) obligatoriness is a sufficient but not a necessary condition for argumenthood; 2) the pre-finite verb position is strictly limited to adjuncts; 3) the "*le faire*" test (a French equivalent of the English "*do so*" test) works only for adjuncts. Other tests (iterativity, movement, etc.) are not categorical and show only tendencies. This situation is all the more less clear for adjectives, see sec. 2.2.

There were few attempts to automatize the distinction. (Merlo and Ferrer, 2006) have implemented a method to automatically distinguish arguments from adjuncts of English verbs and nouns. Their method combines three linguistic diagnoses and approximates them in terms of corpus counts extracted from a manually annotated corpus (Penn-treebank). The tests estimate the following properties: the optionality of the complement, the degree of selection by the head, and the iterativity of a phrase. The availability of lexical resources such as WordNet enable them to use semantic classes that prove to be useful information in the classification task. In the same vein, in (Fabre and Bourigault, 2008) we have designed simple statistical indicators to assess the degree of autonomy of a PP with respect to the verb.

We aim at adopting a similar technique for French adjectives. The task is even more complicated in this case, due to the greater flexibility of the adjective complementation as presented in section 2.2: the optionality of almost all adjective complements, and the variability of the preposition that is used to introduce the complement makes the distinction between complements and adjuncts even more complex.

On the basis of the linguistic properties of adjectives, we propose several statistical measures to evaluate valence properties of candidate frames, focusing in particular on obligatoriness and autonomy of PP (or VPinf) with respect to the adjective. The idea is to consider a PP more likely to be an argument of the adjective if it meets the following three conditions: the adjective is rarely found alone (i.e., without an accompanying frame), the frame is productive (the adjective combines with a large range of nouns or infinitives introduced by the same preposition), and prepositional expansions that are attached to this adjective are mostly introduced by the preposition which appears in the frame. This diagnosis is obtained by computing three measures estimating the optionality and the productivity of the pattern.

1) *Optionality measure*

Since the syntactic realization of adjective complements is usually optional, the information about absence or presence of a complement cannot be considered as a decisive factor. Yet, we can use this information to assess a tendency in the whole corpus. For each adjective, we calculate the proportion of its occurrences that are found without a prepositional expansion. Only 100 out of the 2684 adjectives appear with a preposition in at least 50% of their occurrences, including the adjectives that (Noailly, 1999) has mentioned having an obligatory complement (cf. 2.2).

2) *Productivity measures*

Productivity has mainly been addressed in morphology to estimate the ability of a suffix to produce new words. It can also be used in syntax to assess the regularity of a relation. We use a productivity measure that estimates the regularity of each pattern, by simply counting the

number of different nouns (or infinitives for patterns including a VPinf) that appear as a dependent of the adjective in the corpus. This gives us a first approximation of the behaviour of the pattern. For example, the pattern *applicable* PP[*à*] 'applicable to' is very productive, since 450 different nouns are found in this position, whereas the pattern *difficile* PP[*à*] 'difficult to' is less productive (productivity=29).

Yet, this measure is highly dependent on the frequency of the adjective itself. We then calculate a more precise measure of productivity, the relative productivity, which estimates the part that each preposition takes within the set of all prepositions that can introduce complements of the adjective. Table 4 shows on one example how this is computed.

adjective	pattern	basic productivity	relative productivity = basic productivity / overall productivity
étonnant	de N	35	0.36
	de VINF	17	0.17
	dans N	28	0.29
	pour N	9	0.1
	par N	7	0.07
	(overall productivity = 96)		

Table 4: Productivity of patterns associated with the adjective *étonnant* 'surprising'

Some interesting patterns can exhibit a very low productivity and simultaneously a high relative productivity, such as *croulant sous* 'collapsing under' or *chatouilleux sur* 'touchy about'.

4.2.3. Results

The criteria that we have used so far consist, on one hand, in identifying impersonal constructions (4.3.1) for patterns including a VPinf introduced by *de*, and on the other hand, in assessing argumenthood of PP or VPinf constituents by estimating optionality of the complement and productivity of this relation (4.3.2). These measures allow us to translate traditional linguistic tests of optionality and regularity into very simple statistical cues. This provides a profile of candidate terms (CF) in the corpus. In section 4.3.1 we have shown that the measure adopted to identify impersonal constructions permits to distinguish two sets of candidate frames. If we now turn to the question of argumenthood, we now have indications about the behaviour of each candidate frame in the LM corpus (Table 5). When the three figures are high, this is an indication for the validation of the candidate pattern based on its behaviour in the corpus: the complement is frequently present, the relation is regular and the preposition is frequently associated with the adjective. In Table 5, this is the case for patterns 1, 2, 3, 5. When the three figures are low, as for CF 6, the measures bring about ill-formed patterns. CF 4 and 7 are in-between: for CF 4, the adjective combines very regularly with a PP, but not with the one indicated in the pattern (*sur* 'on') since the CF has a very low productivity. This PP is actually not an argument but an adjunct ('applicable on'). CF 7 is very productive, but the adjective is mostly found alone, without a PP. This result may be interpreted in many ways. In this case, it corresponds to two different meanings of the adjective, *propre* meaning either 'clean' or 'peculiar', and only the second one is strongly associated with the PP[*à*] pattern.

n	candidate frame	Basic productivity	Relative productivity	Proportion of occ. with a PP
1	*âgé*:PP[*de*] aged-of	265	0,91	0,69
2	*aisé*:VPinf[*à*] easy-to	110	0,32	0,16

3	*applicable* :PP[*à*] applicable-to	450	0,74	0,48
4	*applicable* :PP[*sur*] applicable-on	1	0,01	0,48
5	*conforme* :PP[*à*] conform-to	544	0,97	0,68
6	*grave* :PP[*en*] serious-in	3	0,02	0,002
7	*propre* :PP[*à*] peculiar-to	144	0,58	0,005
8	*difficile* :PP[*à*] difficult-to	29	0,01	0,37

Table 5: Profiles of the lexicalised frames

Most CF exhibit high measures of productivity, thus confirming the validity of the data that have been extracted. In order to locate less reliable CF we have set low thresholds, namely:

basic prod. < 10, relative prod. < 0,1, proportion of occ. of the adjective with a PP < 0,1.

24 candidate frames were found with such low values. The examination of these results shows that the method enables us to spot two types of CF:

- CF corresponding to ill-formed patterns that should have been eliminated by the heuristics designed to filter out alternative syntactic constructions (cf. section 3.2.2). For example, the sequence *étroit*:VPinf[*pour*] 'narrow for' is part of the intensifier construction that has not been correctly spotted
- CF corresponding to adjunct relations, such as *dynamic*:PP[*depuis*] 'dynamic since' or *sinistré*:PP[*après*] 'stricken after'. These PPs convey peripheral information about the predicate, such as temporal complements

All 24 CF should be eliminated as they result from errors in the Treebank annotation or in incorrect application of our heuristics. This preliminary result shows that statistical cues provide a useful verification of the candidate frames extracted from the treebank.

So far, we applied the statistical measures to the treebank data. The next step consists in applying them to the LM corpus in order to discover new patterns .

4.3. Extracting new frames

From now on, we will examine new patterns extracted from the LM corpus, i.e., lexicalised frames that have not been found in the treebank. By comparing the overall values of the 3 argumenthood measures (Table 6) we get an idea of the relative quality of patterns. In particular, for the new LM patterns, the average productivity is very low, and so is the average proportion of adjectival occurrences that have prepositional dependents. On the basis of this rough comparison, we may expect that most new patterns will be rejected as subcategorization frames. Our goal is then to identify the minority of genuine valency patterns among these data.

	Treebank Candidate Frames in the LM corpus	*Le Monde* new patterns
proportion of occurrences with a PP	25%	4%
average basic productivity	293	20

average relative productivity	0,5	0,3

Table 6: Comparing Treebank candidate frames and *Le Monde* patterns

4.3.1. Identifying impersonals

Following the method already presented in section 4.2.1., the *impers* measure is tested on 431 new ADJ VPinf[*de*] patterns that are found in the LM corpus. The values that we obtain on this set of patterns are not as clear-cut as what we have observed for Treebank candidate frames. Yet the measure enables us to make a decision for 91,5% of the patterns. Most patterns (84%) get a very high value (more than 80% of the subject forms are potential impersonal constructions), which means that the VPinf must be considered the extraposed subject rather than a complement of the adjective. A few examples are given below:

(23) Il est aberrant, aléatoire, artificiel … de + VPinf
'It is absurd, unpredictable, artificial … to' + VPinf

Only 7% of the patterns get a very low value (less than 20% of the subject forms are potential impersonal constructions), which attests that VPinf is a real complement.

(24) Il est aimable, avide, coupable … de + VPinf
'He is kind, eager, guilty … to' + VPinf

In both cases, when the *impers* measure gets extreme values, it successfully helps to distinguish the two types of constructions. In-between, 8,5% of the patterns get intermediate values and must be more closely looked at. When we examine the contexts in which these 37 patterns occur, two different scenarios emerge. First, they can behave like the *nouveau* VPinf[*de*] pattern that we have mentioned in section 4.2.1: in addition to the impersonal construction we find instances of the nominal subject, corresponding to a specific construction, where the VPinf cannot be considered a dependent of the adjective.

(25) L'erreur serait lourde de le cantonner à cela = Le cantonner à cela serait une lourde erreur
the mistake would-be-serious-to confine him to this
'To confine him to this would be a serious mistake'

In this case, only the impersonal construction indicates a valency element. Second, we have patterns that can match two different constructions: one with a sentential subject (26) and the other with a VPinf complement (27).

(26) Julie est malheureuse d'avoir été exclue
Julie is unhappy to have been excluded

(27) C'est malheureux de parler de ça
It-is-unhappy-to talk about this
'Talking about this is deplorable'

In this case, both patterns must be kept for the adjective.

This measure finally enables us to enhance the lexicon with 362 new lexicalised frames of the form ADJ + [S:subject].

4.3.2. Extracting PP with good valency properties

We now turn to the rest of the data that we have extracted, in order to spot patterns that correspond to new lexicalised frames. The idea is to focus on patterns that exhibit good statistical properties. The difficulty is to turn statistical estimation of valency properties into a binary decision (argument yes/no). Defining thresholds is an arbitrary task, resulting from a compromise between recall and precision. In the present study we have first decided to look at

patterns that exhibit high values for the three measures (basic productivity, relative productivity and optionality), taking as a reference point the average values that we have obtained for the treebank frames (Table 6), namely:

Basic prod.>200, relative prod.>0,5, proportion of occ. of the adjective with a PP >25%.

This first set of values provides very good patterns, but we get very few data. In a second step, we have loosened the constraints in order to get a larger set of data. The new values are:

Basic prod.>5, relative prod.>0,2, proportion of occ. of the adjective with a PP >20%.

The minimum value of basic productivity is considerably reduced following the observation that infrequent patterns can also provide subcategorisation information. On the basis of these statistical counts, we get 199 patterns that allow us to make an initial evaluation of the results. A manual evaluation shows that they are split into 3 categories:

32% are tagging and parsing errors
17% are regular associations but not 'standard' valency information
51% are confirmed lexicalised frames

This first result brings out three preliminary conclusions: the statistical approach must not be considered as a fully automatic procedure, but it should guide an inspection of large amount of data, in order to give priority to information about syntactic dependence that seems to be the most regular and reliable. As a result, 100 more ADJ[PP] or ADJ[VP-inf] frames are thus added to the treebank frames. Second, this experiment shows that the use of large corpora brings to the fore the question of the continuum between arguments and adjuncts. Some PPs may be very regularly associated with adjectives in the corpus without necessarily being syntactic arguments. In the patterns that show regular associations but are not considered 'standard' valency elements, we find information that relates to semantic frames rather than argumenthood:

(28) repérable dans + LOCATION
'that can be spotted in' + LOCATION

(29) perceptible dans + LOCATION
'perceptible in' + LOCATION

(30) indisponible pendant + TIME
'unavailable during + TIME'

The binary decision is all the more difficult to make given that, as seen in section 2.2., traditional linguistic tests do not efficiently distinguish between arguments and adjuncts of adjectives.

Third, the decision of including such patterns in the lexicon depends on how the lexicon is planned to be used. A repertoire of very regular associations that go beyond strict argumenthood can be very useful in the perspective of NLP applications such as the development of parsers which need to resolve PP attachment, or the development of semantic annotation based on semantic frames.

5. Conclusion

We have exposed two complementary methods for detecting adjective valence. Our study shows that the constitution of a valency lexicon can be assisted in different ways by resorting to corpus linguistics methods, combining data coming from a small but linguistically rich and

quite reliable corpus and from a much larger but more noisy corpus. The first corpus provides a small set of good quality frames with the help of linguistic heuristics. The second corpus helps to improve the accuracy of the initial frames by supplying complementary information about syntactic constructions (impersonal vs. personal) and about the regularity of the candidate frames. The large amount of data also provides new patterns equipped with statistical information that guide semi-automatic detection of actual frames.

Regarding the distinction between arguments and adjuncts, the confrontation of frames with data from a large corpus leads us to progress from a binary distinction to a more graduate perception of the regularity of syntactic patterns. Statistical counts show that a PP can be consistently associated with an adjective in the corpus, giving strong indication of selection, without being normally considered a strict argument. This type of information helps to identify in-between cases, as stated by (Manning, 2002) which can be also interesting to describe in terms of colligation or semantic frames.

We plan to further investigate data extracted from the large corpus in order to pursue the identification of frames, but also to examine in more detail dependents of adjectives from the corpus perspective. In particular, this data contains two types of interesting information which should be examined in more detail:

- preposition profiles: prepositions could be ranked according to their average value of relative productivity. For example the complex preposition *à l'égard de* `with respect to', which is not listed as argumental in the PrepLex lexicon (cf. 3.2.2), appears at the top on the list of most productive prepositions
- preposition alternations: corpus data can indicate alternation tendencies of specific prepositions (for example, *devant* `in front of' and *face à* `facing' can be found concurrently with many adjectives).

Such statistical information extracted from a large annotated corpus can benefit not only to the construction of lexical resources for NLP applications but also to descriptive studies of lexico-syntactic properties of adjectives.

6. Notes

[1] Picabia (1978:p.43) mentions two adjectives which appear with an apparent NP complement: *bleu roi* `royal blue' and *rouge cerise* `cherry red'. The two exceptions, however, can be considered multi-word adjectives, cf. (Gross, 1975).

[2] In the LM corpus, no category distinction is made between a preposition introducing a PP and a complementizer preceding a VPinf. According to corpus annotation, both types of phrases are headed by a preposition, hence a type of PP. In this section, the distinction is not made either.

7. References

Abeillé, A., L. Clément and F. Toussenel (2003). "Building a treebank for French". In A. Abeillé (ed.) *Treebanks: Building and using parsed corpora,* 165-187.

Bonami, O. (1999). *Les constructions du verbe: le cas des groupes prépositionnels argumentaux. Analyse syntaxique, sémantique et lexicale*, PhD thesis, University of Paris 7.

Carroll, J. and A. Fang (2004). "The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser", *Proceedings of the 1st International Conference on Natural Language Processing (IJCNL9)*, Sanya City, China, 107-114.

Danlos, L. (1985). *La génération automatique de textes*, Masson.

Fabre, C. and D. Bourigault (2008). ''Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants'', *Journal of French Language Studies*, 18(1), 87-102.

Fort, K. and B. Guillaume (2007). "Preplex: a lexicon of French prepositions for parsing", *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions,* Prague, Czech Republic, 17-24.

Frank, A., L. Sadler, J. van Genabith, J. and A. Way (2002). "From treebank resources to LFG f-structures", In A. Abeillé (ed.) *Treebanks: Building and using parsed corpora,* Kluwer, 367-389.

Gardent, C., B. Guillaume, G. Perrier and I. Falk (2006). "Extraction d'information de sous-catégorisation à partir des tables du LADL", *Proceedings of TALN 2006,* Leuven, 139-148.

Gross, M., (1975). *Méthodes en syntaxe*. Hermann.

Han, C., B. Lavoie, M. Palmer, O. Rambow, R. Kittredge, T. Korelsky and N. Kim (2000). "Handling structural divergences and recovering dropped arguments in a Korean/English machine translation system", *Proceedings of the Association for Machine Translation in the Americas*. Berlin/New York: Springer Verlag, 40-53.

Hornby, A. S. (1989). *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, 4th edition.

Manning C. (2003). ''Probabilistic Syntax''. In R. Bod, J. Hay, and S. Jannedy (eds), *Probabilistic Linguistics*, Cambridge, MA: MIT Press, 289-341.

Mel'cuk, I., N. Arbatchewsky-Jumarie, and A. Clas. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain, Recherches lexico-sémantiques*, vol. I, II, III, IV. Les Presses de l'Université de Montréal.

Merlo, P. and E. Ferrer (2006). "The notion of argument in PP attachment", *Computational Linguistics*, 32(2), 341-378.

Meydan, M. (1999). "La restructuration du GN sujet dans les phrases adjectivales à substantif approprié", *Langages*, 133, 59-80.

Noailly, M. (1999). *L'adjectif en français*, Ophrys.

Paroubek, P., A. Vilnat, I. Robba and C. Ayache (2007). "Les résultats de la campagne EASy d'évaluation des analyseurs syntaxiques du français", *Proceedings of the Workshop EASy, TALN 2007*, Toulouse.

Picabia, L. (1978). *Les constructions adjectivales en français*. Droz, Genève-Paris.

Preiss, J., T. Briscoe and A. Korhonen (2007). “A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora”, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, 912-919.

Procter, P. (ed.) (1978). *Longman Dictionary of Contemporary English*. Longman, Burnt Mill, Harlow.

Surdeanu, M., S. Harabagiu, J. Williams and P. Aarseth (2003). “Using predicate-argument structures for information extraction”, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 8-15